Multilingual Automatic Document Classification Analysis and Translation Phase 1 (MADCAT-P1) Evaluation Plan

1 Introduction

The Multilingual Automatic Document Classification Analysis and Translation (MADCAT) program is a five-year DARPA research program whose purpose is to explore and develop technologies that convert non-English language document images into English transcripts so that the information can be readily used [1]. The goal of the MADCAT evaluation is to measure the performance of these developed technologies. This document describes the evaluation protocols for the first phase of the MADCAT program.

2 Evaluation Tasks

The technologies that the MADCAT program seeks to develop are multidisciplinary. These technologies include document structure extraction, optical character recognition (OCR), and machine translation (MT). Three evaluation tasks are defined for Phase 1 (P1) of the MADCAT program to measure the overall performance of the MADCAT system and to probe the performance of various components within the system. The goal of the evaluation is not only to determine whether the program achieved its target performance goal but also to fully understand the strengths and weaknesses of the system.

To make the evaluation more tractable, segmentation is provided in P1 so that the focus is on measuring the coupling OCR and MT technologies to facilitate Arabic to English translation of images. Segmentation will not be provided in future MADCAT evaluations. The three tasks, summarized in Table 1, are described in detail in the subsections below.

Table 1. Evaluation tasks for MADCAT P1.

Task	Primary	Input	Output	Metric
Document image translation	**Yes**	Arabic document images with manual segmentation	Segmented English translation	HTER
Document image recognition	*No*	Arabic document images with manual segmentation	Segmented Arabic transcription	WER
Document text translation	*No*	Arabic document images with manual segmentation and manual transcription	Segmented English translation	TER

2.1 Document Image Translation

The document image translation is the primary evaluation task. It measures the system performance in translating foreign language document images into accurate and fluent English documents. This task measures the overall performance of the MADCAT system. The input for this task is Arabic document images along with manual segmentation, and the output is to be segmented English translation of these document images.

2.2 Document Image Recognition

The document image recognition is a contrastive component evaluation task. It measures the system performance in recognizing the transcript in the foreign language document images, which is the recognition component of the MADCAT system. The input for this task is Arabic document images along with manual segmentation, and the output is to be segmented Arabic transcription of these document images.

2.3 Document Text Translation

The document text translation is another contrastive component evaluation task. It measures the system performance in translating foreign language document images in accurate and fluent English documents when manual transcription of the source text is given. It measures the translation component of the MADCAT system. The input for this task is Arabic document images along with manual segmentation and manual Arabic transcription, and the output is to be segmented English translation.

3 Data

The data used in P1 of the MADCAT program is drawn from the data used in the Global Autonomous Language Exploitation (GALE) program. Utilizing the same data between the two programs eliminates the domain mismatch allowing the incorporation of MT models developed under GALE for MADCAT use; and because the data properties are well-known, the GALE data provides a controlled data environment. See [5] for details regarding GALE data suitability for MADCAT use.

Two GALE data genres newswire and web text will be used in MADCAT as formal text and informal text, respectively.

3.1 Data Creation for MADCAT

Literate, native Arabic speakers were recruited by the Linguistic Data Consortium (LDC) to act as scribes. The scribes create handwritten copies of the chosen GALE passages according to the agreed distribution of various writing factors. Table 2 lists the target distribution of these writing factors for the MADCAT P1 data.

Table 2. Target distribution of various writing factors for the data used in MADCAT.

Writing instrument	Writing Surface	Writing Speed
90% ballpoint pen	75% unlined white paper	90% normal
10% pencil	25% lined paper	5% fast
		5% careful

3.2 Data Sets

3.2.1 Formal Evaluation Data

The formal evaluation data will be chosen from the GALE Phase 3 (P3) evaluation data pool. The data will be selected such that it maximizes the overlap with GALE P3 evaluation data set to take advantage of the time reduction in producing the transcription and translation on the data. The overlap also will enable direct comparisons with the GALE results.

Two MT systems installed at the LDC will be used to produce MT for the GALE P3 Arabic evaluation data pool and the GALE P2 Arabic evaluation set. TER scores (see section 7.2) will be computed for the MT of the two sets. A subset of the P3 pool will be selected so that its TER distribution matches the TER distribution of the GALE P2 set.

Table 3. Target size for MADCAT P1 formal evaluation data set.

Genre	Newswire	Web Text
Source	GALE P3 Eval	GALE P3 Eval
Number of pages	160	160
Arabic tokens per page	125	125
Number of scribe per page	6	6
Number of unique scribes	50 (25 from training, 25 new)	

3.2.2 Pilot Evaluation Data

A MADCAT pilot evaluation will occur in early Fall to test the evaluation protocols. The pilot evaluation data set will be drawn from the formal P1 evaluation data set and will consist of about one-half the number of pages. The pilot evaluation data will contain two of the six scribes. Table 4 lists the target size for the pilot evaluation set.

Table 4. Target size for MADCAT P1 pilot evaluation data set.

Genre	Newswire	Web Text
Source	GALE P3 Eval	GALE P3 Eval
Number of pages	80	80
Arabic tokens per page	125	125
Number of scribe per page	2	2
Number of unique scribes	24 (14 from training, 10 new)	

3.2.3 Development Data

The MADCAT P1 development set will come from the GALE P1 and GALE un-sequestered P2 evaluation data sets. No special selection procedures were applied. Table 5 lists the target size for the development set.

Table 5. Target size for MADCAT development data set.

Genre	Newswire	Web Text
Source	GALE P1-P2 Eval	GALE P1-P2 Eval
Number of pages	160	160
Arabic tokens per page	125	125
Number of scribe per page	2	2
Number of unique scribes	50 (25 from training, 25 new)	

3.2.4 Training Data

The MADCAT P1 training data will come from a subset of the GALE P1-P3 parallel text training data releases. It contains a mix of newswire and web text with five scribes per page with 100 unique scribes. Table 6 lists the target size for the training set.

Table 6. Target size for MADCAT training data set.

Genre	Newswire/Web Text
Source	GALE P1-P3 parallel text
Number of pages	2000
Arabic tokens per page	125
Number of scribe per page	5
Number of unique scribes	100

4 Evaluation Rules and Restrictions

The following list of rules and restrictions must be observed:

- Each page is to be processed independently.
- Adaptation across multiple pages is not allowed.
- Interaction with the evaluation test data before submission of system results is not allowed. This includes both human interaction and automatic probing of the data.

5 Data File Format

All data created for MADCAT P1 is stored in an XML format that defines storage elements that capture the various annotation layers in a document image. The format is defined in detail in version v4h2 of the MADCAT Format Specifications document [2].

5.1 Reference Data

Each reference file contains two main layers of information. The first layer contains the word and line level segmentation for the document image, and the second layer contains the transcription and translation of the text in the image. See section 3 of [2]. The reference files are identified with the extension "`.madcat.xml`".

For example: `<BASENAME>.madcat.xml`

5.2 Input Data

Each input file is derived from the corresponding reference file, and depending on the evaluation task, certain information is removed from the reference file and used as input to the MADCAT system.

For the document image translation and document image transcription tasks, information from the translation and transcription layers is removed. These input files are identified with the "`.seg.madcat.xml`" extension.

For example: `<BASENAME>.seg.madcat.xml`

For the document text translation task, information from the translation layer is removed. These input files are identified with the "`.textseg.madcat.xml`" extension.

For example: `<BASENAME>.textseg.madcat.xml`

5.3 Output Data

Depending on the input, each system output file is to contain the translation and/or transcription information as produced by the MADCAT system.

For the document image translation and document image transcription tasks, the MADCAT system is to output the transcription and translation information. These output files are identified with the "`.sys.madcat.xml`" extension.

For example: `<BASENAME>.sys.madcat.xml`

For the document text translation task, the MADCAT system is to output the translation information. These output files are identified with the "`transys.madcat.xml`" extension.

For example: `<BASENAME>.transys.madcat.xml`

6 Post-Editing Protocol

Each system output will be edited for correctness by two independent teams of editors. A team consists of a pair of editors with one editor making edits in a first pass and a second editor acting as a reviewer. The reviewer checks the first pass edits for correctness while making additional modifications if needed. The output of the reviewer is the final version from the team.

The output of the two reviewers is compared at the segment level, choosing the segment that has the lower HTER score (see section 7.2). The final document level HTER score is the resulting HTER when choosing the lower segment across the both sets of post-edited MT.

6.1 Post-Editing Kit and Editor Team Assignment

NIST defines "kits" as a collection of system output to be post edited by two post editing teams.

Each kit is to have around 600 words, which was determined to be a reasonable amount of data for an average editor to edit in one session. Each kit contains a mixture of the two genres, newswire and web text, an attempt to reduce a specific editor team effect on a single genre.

Each kit is assigned to two teams of editors. There are 15 teams of editors. To the extent possible kits are assigned to editor pairs as to maximize the overlap for comparing editor team statistics.

7 Evaluation Metrics

The three evaluation tasks described in section 2 are measured by the three metrics described in the subsections below.

7.1 TER

The system performance on the document text translation task is measured by TER [3]. Short for Translation Edit Rate, TER is an edit distance metric that measures translation quality. It calculates the exact match between the system translation and the reference translation.

$$TER = \frac{(\#insertions + \#deletions + \#substitutions + \#shifts)}{\#reference\_translated\_words}$$

7.2 HTER

The system performance on the document image translation task is measured by HTER. Short for Human-mediated Translation Edit Rate, HTER is a modified version of TER where it involves a human editor modifying the system output such that it contains the exact meaning of the reference translation. It is the primary metric of translation quality for MADCAT.

7.3 WER

The system performance on the document image transcription task is measured by WER. Short for Word Error Rate, WER is an edit distance metric that measures transcription quality. It is defined as the minimum number of steps taken to transform the system transcript to have the exact words as the reference transcript.

$$WER = \frac{(\#insertions + \#deletions + \#substitutions)}{\#reference\_transcribed\_words}$$

8 Scoring Package

NIST is developing a scoring package to facilitate the calculation of the above metrics. The package utilizes the software tercom-0.7.25 developed by UMD-BBN [3] as well as those developed internally at NIST [4].

Normalization is to be performed on the system output prior to scoring. For the translation tasks, punctuations in the reference and system translations are tokenized. In addition, the scoring preserves case. For the transcription task, if any diacritic information is present in the reference and system transcripts, it is removed. Transcription scoring also takes case into account.

Segments containing scribe errors are to be included as-is for post editing. A stand-off annotation file will identify all segments that contain scribe errors so that such segments may be analyzed separately.

9 Submission of Results

Submission of the evaluation results will be done via FTP:

- Create a directory where the system output will reside
- Place the output in that directory
- Tar and compress the directory
- FTP the tar and compressed file to jaguar.ncsl.nist.gov/madcat/incoming
- Send an email to madcat_poc@nist.gov to notify the submission was made

For example:

- `mkdir plato_1`
- `cp *.{sys|transys}.madcat.xml plato_1`
- `tar zcvf plato_1.tgz plato_1`
- `ftp jaguar.ncsl.nist.gov (anonymous login with email as password)`
    - `binary`
    - `cd incoming`
    - `put plato_1.tgz`
    - `bye`
- `send an email to madcat_poc@nist.gov`

10 Schedule

Training & Development Data	
Training data release 1	June 3, 2008
Training data release 2	July 5, 2008
Training data release 3	July 31, 2008
Training data release 4	August 29, 2008
Development data release	September 9, 2008
Pilot Evaluation	
Pilot evaluation data release by LDC (segmentation only)	September 23, 2008
Pilot evaluation results due to NIST	October 7, 2008
Pilot evaluation PE (post-editing) starts	October 16, 2008
Pilot evaluation data release by NIST (segmentation+ transcription)	October 21, 2008
Pilot evaluation PE ends	November 12, 2008
Final results to DARPA	November 14, 2008
Formal Evaluation	
Formal evaluation data release	January 20, 2009
Formal evaluation results due to NIST	February 20, 2009
Formal evaluation PE starts	February 20, 2009
Formal evaluation PE ends	April 17, 2009
Final results to DARPA	May 6, 2009
Miscellaneous	
Scoring software release	July 31, 2008

11 Glossary of Terms

Document – a naturally occurring unit of original source data of variable length collected by LDC

Passage – a sub-section within a document chosen for evaluation

Manuscript – a copy of a passage created by a scribe

Page – one of the leaves in a manuscript created by a scribe. This is the basic unit of evaluation

Scribe – a person who creates a handwritten copy of one or more passages

12 References

[1] J. Olive, “Multilingual Automatic Document Classification Analysis and Translation (MADCAT) SOL BAA 07-38 Proposer Information Pamplet”, DARPA/IPTO, 2007.

[2] MADCATDataFormatSpec_V4h2.zip at https://madcatwiki.ldc.upenn.edu/madcatwiki/index.php/Data_Format

[3] M. Snover, B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," *Proceedings of Association for Machine Translation in the Americas*, 2006.

[4] J. Fiscus, J. Ajot, N. Radde, and C. Laprun, “Multiple Dimension Levenshtein Edit Distance Calculations for Evaluating Automatic Speech Recognition Systems During Simultaneous Speech", *Proceedings of LREC*, 2006.

[5] MADCAT_Data_Planning_4_Feb_2008v11.ppt at https://madcatwiki.ldc.upenn.edu/madcatwiki/index.php/Meetings/Phase1/DARPABrief